Claims

1. A method for monitoring abnormalities in a data stream, comprising the steps of:

receiving a plurality of objects in the data stream;

creating one or more clusters from the plurality of objects, wherein at least a portion of the one or more clusters comprise statistical data of the respective cluster; and

determining from the statistical data whether one or more abnormalities exist in the data stream.

2. The method of claim 1, wherein the step of creating one or more clusters further comprises:

computing one or more similarity values for a given object relating to one or more existing clusters; and

determining a closest cluster for the object based on the one or more similarity values.

3. The method of claim 2, further comprising the steps of:

determining whether to add the object to the closest cluster;

adding the object to the closest cluster when determined and updating the statistical data of the closest cluster; and

creating a new cluster comprising the object when the object is not added to the closest cluster, and generating statistical data of the new cluster.

4. The method of claim 3, wherein the step of determining whether to add the object to the closest cluster further comprises the step of determining if the similarity value is greater than a user-defined threshold.

5. The method of claim 1, wherein the step of determining from the statistical data whether one or more abnormalities exist further comprises the steps of:

determining which clusters present at a first time were not present at a second time, wherein the second time is before the first time;

determining which of the clusters, present at the first time and not present at the second time, contain fewer than a user-defined number of objects; and

reporting clusters with fewer than the user-defined number of objects as abnormalities.

6. The method of claim 1, wherein the statistical data of each cluster is stored using an incremental updating process.

7. The method of claim 1, wherein the statistical data of each cluster comprises one or more statistical counts of each pairwise attribute.

8. The method of claim 1, wherein the statistical data of each cluster comprises one or more statistical counts of each categorical attribute.

9. The method of claim 1, wherein the statistical data of each cluster comprises a number of objects in each cluster.

10. The method of claim 1, wherein the statistical data is stored periodically at intervals chosen based on a pyramidal distribution.

11. The method of claim 1, wherein the step of creating one or more clusters further comprises the step of applying one or more weights to one or more attributes.

12. The method of claim 1, wherein abnormalities comprise intrusions in a network.

13. The method of claim 12, wherein the step of receiving a plurality of objects further comprises the step of collecting source IP (Internet Protocol) address data, destination IP address data and signature data.

14. The method of claim 12, wherein the step of creating one or more clusters further comprises the step of clustering source IP address data, destination IP address data and signature data.

15. The method of claim 12, wherein the step of determining from the statistical data whether one or more abnormalities exist further comprises the step of detecting one or more intrusions from statistical data of source IP address data, destination IP address data and signature data.

16. Apparatus for monitoring abnormalities in a data stream, comprising:

a memory; and

at least one processor coupled to the memory and operative to: (i) receive a plurality of objects in the data stream; (ii) create one or more clusters from the plurality of objects, wherein at least a portion of the one or more clusters comprise statistical data of the respective cluster; and (iii) determine from the statistical data whether one or more abnormalities exist in the data stream.

17. The apparatus of claim 16, wherein the operation of creating one or more clusters further comprises:

computing one or more similarity values for a given object relating to one or more existing clusters; and

determining a closest cluster for the object based on the one or more similarity values.

18. The apparatus of claim 17, further comprising:

determining whether to add the object to the closest cluster;

adding the object to the closest cluster when determined and updating the statistical data of the closest cluster; and

creating a new cluster comprising the object when the object is not added to the closest cluster, and generating statistical data of the new cluster.

19. The apparatus of claim 18, wherein determining whether to add the object to the closest cluster further comprises determining if the similarity value is greater than a user defined threshold.

20. The apparatus of claim 17, wherein the operation of determining from the statistical data whether one or more abnormalities exist further comprises:

determining which clusters present at a first time were not present at a second time, wherein the second time is before the first time;

determining which of the clusters, present at the first time and not present at the second time, contain fewer than a user defined number of objects; and

reporting clusters with fewer than a defined number of objects as abnormalities.

21. The apparatus of claim 16, wherein the statistical data of each cluster is stored using an incremental updating process.

22. The apparatus of claim 16, wherein the statistical data of each cluster comprises one or more statistical counts of each pairwise attribute.

23. The apparatus of claim 16, wherein the statistical data of each cluster comprises one or more statistical counts of each categorical attribute.

24. The apparatus of claim 16, wherein the statistical data of each cluster comprises a number of objects in each cluster.

25. The apparatus of claim 16, wherein the statistical data is stored periodically at intervals chosen based on a pyramidal distribution.

26. The apparatus of claim 16, wherein the operation of creating one or more clusters further comprises applying one or more weights to one or more attributes.

27. The apparatus of claim 16, wherein abnormalities comprise intrusions in a network.

28. The apparatus of claim 27, wherein the operation of receiving a plurality of objects further comprises collecting source IP address data, destination IP address data and signature data.

29. The apparatus of claim 27, wherein the operation of creating one or more clusters further comprises clustering source IP address data, destination IP address data and signature data.

30. The apparatus of claim 27, wherein the operation of determining from the statistical data whether one or more abnormalities exist further comprises detecting one or more intrusions from statistical data of source IP address data, destination IP address data, and signature data.

31. An article of manufacture for monitoring abnormalities in a data stream, comprising a machine readable medium containing one or more programs which when executed implement the steps of:

receiving a plurality of objects in the data stream;

creating one or more clusters from the plurality of objects, wherein at least a portion of the one or more clusters comprise statistical data of the respective cluster; and

determining from the statistical data whether one or more abnormalities exist in the data stream.